

Model-Based Motion Filtering for Improving Arm Gesture Recognition Performance

Greg S. Schmidt¹ and Donald H. House²

¹ ITT Industries at the Virtual Reality Laboratory, Naval Research Laboratory
Washington, DC, USA, schmidt@ait.nrl.navy.mil

² Visualization Laboratory, College of Architecture, Texas A&M University
College Station, TX, USA, house@viz.tamu.edu

Abstract. We describe a model-based motion filtering process that, when applied to human arm motion data, leads to improved arm gesture recognition. Arm movements can be viewed as responses to muscle actuations that are guided by responses of the nervous system. Our motion filtering method makes strides towards capturing this structure by integrating a dynamic model with a control system for the arm. We hypothesize that embedding human performance knowledge into the processing of arm movements will lead to better recognition performance. We present details for the design of our filter, our evaluation of the filter from both expert-user and multiple-user pilot studies. Our results show that the filter has a positive impact on recognition performance for arm gestures.

1. Introduction

Gesture recognition techniques have been studied extensively in recent years because of their potential for application in user interfaces. It has long been a goal to apply the “natural” communication means that humans employ with each other to the interfaces of computers. People commonly use arm and hand gestures, ranging from simple actions of “pointing” to more complex gestures that express their feelings and enhance communication. Having the ability to recognize arm gestures by computer would create many possibilities to improve application interfaces, especially those requiring difficult data manipulations (e.g., 3D transformations). Pointing operations would certainly be an effective means to infer directional information such as where to move an object in a computer environment. To date no method has been found for arm gesture recognition that is both very accurate and extendable to broad sets of gestures. Typical approaches (e.g., HMMs, neural networks) have focused on applying analytical methods for breaking down motion sequences and recognizing patterns.

The human model-based approach takes into consideration that while a person is making gestures, the resulting motions and poses are played out by a known, rather than an unknown, process. The gestures can be viewed as responses of a skeletal frame to muscle actuations that are made in response to control signals originating in the nervous system. The structure of the skeleton, joints, and musculature, is well known and well studied. The neural control systems that actuate the muscles are becoming better understood. With a solid

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 2004		2. REPORT TYPE		3. DATES COVERED 00-00-2004 to 00-00-2004	
4. TITLE AND SUBTITLE Model-Based Motion Filtering for Improving Arm Gesture Recognition Performance				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) ITT Industries at the Virtual Reality Laboratory,Naval Research Laboratory,Washington,DC,20375				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES International Gesture Workshop 2003, Lecture Notes in Computer Science, Vol. 2915, Springer-Verlag 2004					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 20	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

model of human dynamics and control, much of the analytical heuristic guess-work might be eliminated. The arm is a good subject for testing model-based approaches because it is an articulated structure with well understood musculature and fairly large inertias that must have a significant effect on gesture performance.

We have designed a motion adaptation filter for enhancing the signal leading to the gesture recognizer that integrates both physical and control models of human gesture. Our technique uses two motion filters: one augmented with a “learned” parametric gesture sequence and control system, and the other unaugmented. Our method for incorporating process knowledge—the model and its dynamics—is the extended Kalman filter, though any process estimation filter could be used that can handle non-linearities. The squared difference between the outputs of both filters is computed and normalized, giving a score that can be used by the recognition system.

Our working hypothesis is that the motion adaptation filter will improve the unknown signal’s quality enough to improve or simplify the recognition process. We tested the hypothesis by integrating the filter with a simple template gesture recognition system, although our filter can be integrated with any standard type of gesture recognition system. To determine the impact that our filter has on arm-movement recognition performance, we tested the system with an expert user performing multiple sets of gestures and with a multiple-user pilot study.

2. Related Work

Here we briefly describe the most common recognition methods and previous related work utilizing human model-based approaches. More complete details can be found in surveys by Watson [1], Aggarwal and Cai [2], Pavlovic et al. [3] and our technical report [4].

2.1. Overview of Recognition Methodologies

The common methodologies that have been used for motion and gesture recognition are: (1) template matching [1], (2) feature-based [1], (3) statistical [5], [6] and (4) multimodal probabilistic combination [7]. By far the most popular recognition methods are feature based neural networks (e.g., [8], [9], [10]) and statistical hidden Markov models (HMMs) (e.g., [11], [12], [13]). Each approach has drawbacks that either affect performance or limit usability. One of the major drawbacks is that most depend on user-specific training and parameter tuning.

The template approach compares the unclassified input sequence with a set of predefined template patterns. The algorithm requires preliminary work to generate a set of gesture patterns, and usually has poor performance due to the difficulty of spatially and temporally aligning the input with the template patterns [1].

The neural network approach typically uses a pre-determined set of common discriminating features, estimates covariances during a training process, and uses a discriminator (e.g., the classic linear discriminator [14]) to classify gestures.

The drawback of this method is that features are manually selected and time-consuming training is involved [1].

The HMM method is a variant of a finite state machine characterized by a set of states, a set of observation symbols for each state, and probability distributions for state transitions, observation symbols and initial states [5]. The major drawbacks of the HMMs are: (1) they require a set of training gestures to generate the state transition network and tune parameters; (2) they make the assumption that successive observed operations are independent, which is typically not the case with human motion [15].

In a multimodal recognition process, two or more human senses are captured and/or two or more capturing technologies are combined. The multiple inputs are processed by a classifier, which rates the set of possible output patterns with a value based upon the likelihood of a match. The set of probabilities for each input are then combined in a manner to be able to select the most likely pattern. Many groups have explored combining speech and gesture (e.g., Cohen et al. [7], Vo and Waibel [16]).

2.2. Methods Utilizing Human Model-Based Approaches

Human model-based approaches integrate a model of human motion, typically approximated as a dynamic process and control system, into the process of filtering motion capture data of human movements. Such a model-based approach seems to have first appeared in Pentland and Horowitz [17]. Model-based approaches to motion generation for animation have been utilized by Zordan and Hodgins [18], Metaxas [19] and others. Wren and Pentland [20] applied dynamics to a 3D skeletal model for a tracking application. They applied 2D measurements from image features and combined them with the extended Kalman filter to drive the 3D model. Their resulting tracking system was able to tolerate temporary image occlusions and the presence of multiple people in the tracked area. In more recent work [21] they explored the notion that people utilize muscles to actively shape purposeful motion. In earlier work [22], we explored the use of a simple particle model for arm motion recognition performance.

3. Background

Here we give the background for methods that we utilized and integrated in the design of our filter.

3.1. Extended Kalman Filter

The extended Kalman filter (EKF) [23] estimates both the time sequence of states of an input data stream and a statistical model of that data stream. The EKF differs from the standard Kalman filter [24] in that it can be used to estimate a process that is non-linear and/or handle a measurement relationship to the process that is non-linear. The EKF can be augmented by a dynamic model of the system being tracked, and knowledge of the reliability of this model. Simply described, the filter is a set of time update equations that estimate the next state vector, current error covariance and the Kalman gain. The Kalman

gain affects the weighting of measurement data versus the control model in determining the next state vector estimate. If the dynamic model is left out or is unreliable, the Kalman gain is high and the filter simply smoothes the input data.

The EKF's prediction equations may be written

$$\begin{aligned}\mathbf{x}_{i+1}^- &= f(\mathbf{x}_i, \mathbf{u}_i, 0) \\ P_{i+1}^- &= A_i P_i A_i^T + W_i Q_i W_i^T,\end{aligned}\tag{1}$$

where f estimates the *a priori* state vector \mathbf{x}_{i+1}^- , as a function of the current state vector \mathbf{x}_i , and the process model vector \mathbf{u}_i at the current time step. P_i and P_{i+1}^- are the current and *a priori* estimated error covariances, Q_i is the process model error covariance, A and W are the Jacobians of f with respect to the state \mathbf{x} and a vector of random variables \mathbf{w} .

The filter's update equations may be written

$$\begin{aligned}K_i &= P_i^- H_i^T (H_i P_i^- H_i^T + V_i R_i V_i^T)^{-1} \\ \mathbf{x}_i &= \mathbf{x}_i^- + K_i (\mathbf{z}_i - h(\mathbf{x}_i^-, 0)) \\ P_i &= (I - K_i H_i) P_i^-, \end{aligned}\tag{2}$$

where K_i is the current Kalman gain matrix, \mathbf{v} is a vector of random variables, h relates the state vector to the measurement vector \mathbf{z}_i , R_i is the measurement error covariance, and H and V are the Jacobians of h with respect to \mathbf{x} and \mathbf{v} .

3.2. Lagrangian Formulation for Dynamics

The Lagrangian formulation for dynamics is particularly appropriate for articulated systems. The Lagrangian

$$L(\mathbf{q}, \dot{\mathbf{q}}) = E_k(\mathbf{q}, \dot{\mathbf{q}}) - E_p(\mathbf{q})\tag{3}$$

is the difference between the kinetic energy E_k and potential energy E_p of the system as a function of state \mathbf{q} . The state is a set of generalized joint coordinates and its rate $\dot{\mathbf{q}}$ is a set of related velocities. The Lagrangian formulation for the dynamics of a system is

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = \tau_i, \quad i = 1, \dots, m,\tag{4}$$

where τ is the set of externally applied or nonconservative forces and torques [25].

Solutions to Equation 4 can be found in closed form, which are more efficient and readily parameterizable than the open form derivations generated by the Featherstone algorithm [26], which is a very efficient rendition of the Newton-Euler approach to dynamics [27]. On the other hand, the open form derivations do have the advantage that they can be easily extended to handle large sets of joint-space configurations.

4. Motion Adaptation Filter

The design of our model-based motion adaptation filter is shown in Figure 1. Its two extended Kalman filters each contain a model of the human arm and its dynamics. Only one is augmented with a model of a control system acting

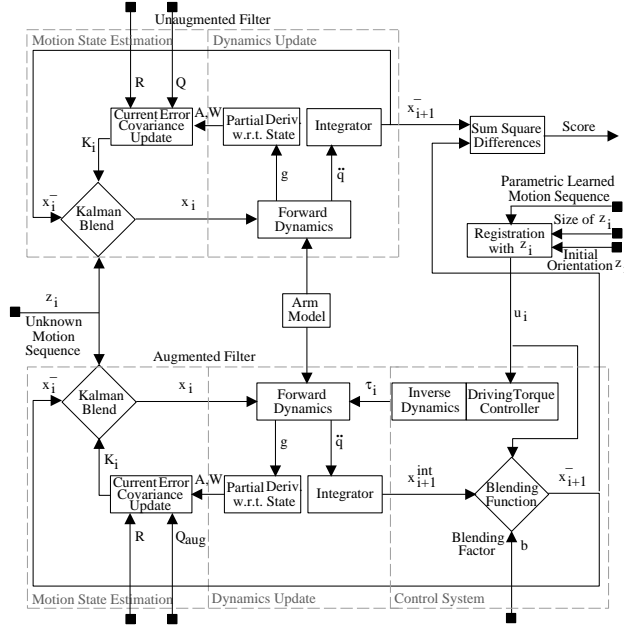


Fig. 1. Motion Adaptation Filter

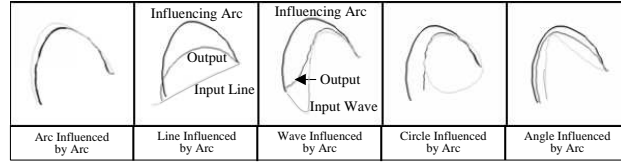


Fig. 2. Five Gestures Influenced by an Arc Motion Sequence

on the arm. The input unknown motion sequence is passed through each filter, compared and a score is computed, which is used as output for the motion adaptation filter.

The unaugmented filter simply smoothes the input motion sequence. Since it contains a control system, the augmented filter attempts to influence the raw input motion sequence to follow a learned motion sequence. We illustrate this notion in Figure 2 by showing five different motion sequences (arc, line, wave, circle and angle) as influenced by a control system generating an arc. Each sequence starts on the right side and proceeds towards the left. The darkest grey line indicates the “influencing” arc sequence, the lightest grey is the input sequence, and the mid-grey is the output sequence. The images show the degree of influence that the arc controller has on each of the input sequences. The degree of this influence is determined by the Kalman gain.

The unaugmented and augmented filters both contain units for motion state estimation and dynamics update. The state estimation unit blends the input motion sequence with the current state vector and passes the data to the dy-

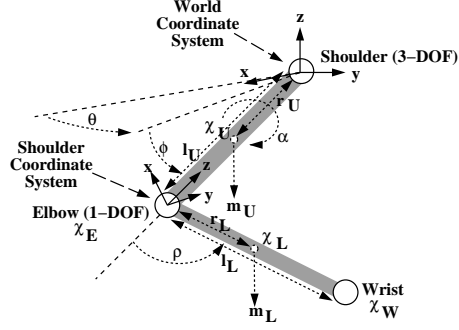


Fig. 3. Articulated Arm Model

namics update process. There, forward dynamics are performed on the state vector producing angular accelerations. These are numerically integrated generating the next state vector. The next state vector is fed back into the system at the Kalman blend and sent to be compared with the output from the augmented filter. The Kalman gain is updated from the current error covariance which is subsequently updated by data from the dynamics update process.

The augmented filter's control system is composed of a driving torque controller and a blending function. Torques used by the controller are derived from the parametric learned motion sequence and model and applied to the forward dynamics of the system. After numerical integration, an intermediate state vector is passed to the blending function where it is mixed with the aligned and parameterized learned motion sequence producing the next state vector. The motivation behind the augmented filter is that if the input motion sequence matches closely to the learned motion sequence (e.g., in Figure 2 the arc in arc module), then the resulting trajectory should be very similar to the input. Thus the trajectories output by the unaugmented and augmented filters will be nearly identical, and the output score will be small. However, if the input motion sequence is dissimilar (e.g., in Figure 2 the line in arc module) to the learned sequence, the trajectories will differ greatly and likewise the score will be large.

4.1. Arm Model

A dynamic articulated model of a human arm is integrated into the filter. The arm model consists of a 3-DOF shoulder joint, a 1-DOF elbow joint and cylinder linkages between the shoulder and elbow, and between the elbow and wrist. The model is shown in Figure 3. We ignore the wrist twist in the lower arm. We also capture the three degrees of freedom for the torso, which is used to produce a relative coordinate system for the arm. The three degrees of freedom from the torso are eliminated after the coordinate transformation takes place between the torso and shoulder.

The position of the wrist and elbow can be determined by using the kinematics equations of motion for the arm model. The equations are parameterized using joint angles for each degree of freedom of the joints in the model. They are

$$\begin{aligned} \chi_E &= (-l_U S_\theta C_\phi, -l_U S_\theta S_\phi, -l_U C_\theta)^T, \\ \chi_W &= R_z(\phi) R_y(\theta) (-l_L S_\rho C_\alpha, -l_L S_\rho S_\alpha, -l_L C_\rho)^T, \end{aligned} \quad (5)$$

where χ_E and χ_W are the positions of the elbow and wrist, respectively, l_U and l_L are the corresponding lengths of the upper and lower arm, $R_z(\phi)$ and $R_y(\theta)$ are rotation matrices about the respective axes z and y , and S and C are sines and cosines of angles of rotation θ , ϕ , α and ρ .

4.2. Motion State Estimation

Motion state estimation is used to predict the state vector at the next time step for the current state of measured input, dynamic model and statistical models of the measured and control systems. The statistics for the measurement process and control system are in the form of error covariance matrices and are pre-determined using training and measurements from the user workspace. They are used by the EKF along with data from the dynamics update process to determine the current Kalman gain.

The Kalman gain is critical for state estimation in the system and requires knowledge from the dynamics and measurement processes. These data include the four (8x8)-Jacobian matrices A , W , H and V from Equations 1 and 2, which relate the process and measurement system's state vectors to the current state vector. The analytic equations for the elements of these matrices are predetermined and their values updated as the filter operates. They are

$$A = \begin{bmatrix} \frac{1}{t} \frac{\partial g}{\partial \mathbf{q}} & 1 + t \frac{\partial g}{\partial \dot{\mathbf{q}}} \end{bmatrix}, \quad W = \begin{bmatrix} \frac{1}{t} \frac{\partial g}{\partial \mathbf{w}_1} & 1 + t \frac{\partial g}{\partial \mathbf{w}_2} \end{bmatrix},$$

and $H = V = \mathbf{I}$ where \mathbf{I} is the 8x8-identity matrix. The matrices A and W are updated by taking the partial derivatives with respect to the current state vector of their respective complete forward dynamics equation g . The augmented and unaugmented filters have different formulations. The formulation for the augmented filter is

$$g(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{w}_1, \mathbf{w}_2) = B'^{-1} \left[\frac{1}{2} (\dot{\mathbf{q}} + \mathbf{w}_2)^T \frac{\partial}{\partial \dot{\mathbf{q}}} [B'] (\dot{\mathbf{q}} + \mathbf{w}_2) - \dot{B}' (\dot{\mathbf{q}} + \mathbf{w}_2) + \tau(\mathbf{q}^m, \dot{\mathbf{q}}^m) \right], \quad (6)$$

and for the unaugmented filter is

$$g(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{w}_1, \mathbf{w}_2) = B'^{-1} \left[\frac{1}{2} (\dot{\mathbf{q}} + \mathbf{w}_2)^T \frac{\partial}{\partial \dot{\mathbf{q}}} [B'] (\dot{\mathbf{q}} + \mathbf{w}_2) - \dot{B}' (\dot{\mathbf{q}} + \mathbf{w}_2) \right], \quad (7)$$

where \mathbf{w}_1 and \mathbf{w}_2 are vectors of random variables representing “white” noise with zero mean and constant variance associated with the process model's state vector and velocities, respectively. B and \dot{B} are the inertia matrices defined in Section 4.3 composed of members from the state vector \mathbf{q} and angular velocities $\dot{\mathbf{q}}$. B' and \dot{B}' are similar matrices to B and \dot{B} but wherever an element of \mathbf{q} and $\dot{\mathbf{q}}$ appears, the appropriate random variable from the vectors \mathbf{w}_1 or \mathbf{w}_2 is added to that member. For example, if θ appears in an element of matrix B , then in B' it is replaced by $\theta + w_{11}$, the first element in the vector \mathbf{w}_1 , since θ is the first element in \mathbf{q} .

4.3. Dynamics Update

The dynamics update process provides parameter updates for motion state estimation and the control system. It takes the current state of the system and

the arm model (and a set of torques for the augmented filter), and performs forward dynamics to produce the parameter update functions g (described in Section 4.2) and the angular accelerations \ddot{q} . Our experiments showed that Euler numerical integration [28] was adequate for updating the next state vector using the accelerations.

The forward dynamics equation for the 4-DOF articulated arm model generates the angular accelerations and is used to derive the complete forward dynamics equations (Equations 6 and 7). In order to derive these equations, the masses, lengths and moments of inertia of the arm segments are needed. Each arm segment is represented by a thin cylinder rotating about its endpoint. The center of mass for each cylinder is estimated using data from a study on anthropometric parameters for the human body in [29]. The data gives estimations for the segmental center of mass (COM) locations expressed in percentages of the segment lengths. These are measured from the proximal end of the segments. The moment of inertia for each segment is computed by combining the inertia tensor of the representative cylinder body and inertial component associated with the shift of its COM to the endpoint. The inertial components associated with the shift of the COM are

$$\begin{aligned}\chi_U &= (-r_U S_\theta C_\phi, -r_U S_\theta S_\phi, -r_U C_\theta)^T, \\ \chi_L &= R_z(\phi)R_y(\theta)(-r_L S_\rho C_\alpha, -r_L S_\rho S_\alpha, -r_L C_\rho)^T,\end{aligned}\quad (8)$$

where χ_U and χ_L are the positions in Cartesian world space of the estimated COMs of the upper and lower arm, respectively, and r_U and r_L are the corresponding radial distances from the shoulder and elbow, respectively. Time derivatives are taken to get the angular velocities at the estimated COMs of the arm segments. These are

$$\dot{\chi}_i = J_i \dot{q}, \quad i = \{U, L\} \quad (9)$$

where the Jacobian matrices $J_U = \frac{\partial \chi_U}{\partial q}$ and $J_L = \frac{\partial \chi_L}{\partial q}$, and $\dot{q} = (\dot{\theta}, \dot{\phi}, \dot{\alpha}, \dot{\rho})^T$. The inertial components are

$$\begin{aligned}I_U &= m_U J_U^T J_U + I_{body_U}, \\ I_L &= m_L J_L^T J_L + I_{body_L},\end{aligned}\quad (10)$$

where I_U and I_L are the inertial components of the upper and lower arm, respectively, m_U and m_L are the estimated masses of the arm segments, and I_{body_U} and I_{body_L} are diagonal matrices representing the thin cylinder body inertias about each parameterized of the axes θ , ϕ , α and ρ . The elements in I_{body_U} and I_{body_L} are determined by converting the cylinder's Euclidean coordinates to spherical coordinates.

The angular velocities and inertias are used to compute the kinetic energy

$$E_k = \frac{1}{2} \dot{q}^T B \dot{q}, \quad (11)$$

where $B = I_U + I_L$. The potential energy is given as

$$\begin{aligned}E_p &= -m_U g r_U C_t \\ &\quad -m_L g [l_U C_t - r_L S_r C_a S_t + r_L C_t C_r],\end{aligned}\quad (12)$$

where g is the gravitational constant. The two energy terms are used for the Lagrangian, L , of Equations 3 and 4. The dynamics equations are computed and solved for angular acceleration

$$\ddot{q} = B^{-1}[\frac{1}{2}\dot{q}^T \frac{\partial}{\partial \dot{q}}[B]\dot{q} - \dot{B}\dot{q} + \tau], \quad (13)$$

where τ is the set of applied torques.

4.4. Control System

Our control system acts as an analogue to the motor nervous system in the human body, influencing how the learned motion sequence acts on the current motion state. It is composed of a driving torque controller and a blending function. The driving torque controller uses data from the learned motion sequence and arm model and performs inverse dynamics, which generates torques for the dynamics update process. The blending function combines the learned motion sequence with an intermediate state vector from the dynamics update process. The degree of its influence is controlled by a fixed predetermined blending factor. The learned motion sequence also remains fixed throughout the iteration of the filter. We see the driving torque controller as analogous to an open-loop predictive control and the blending function as analogous to proprioceptive and sensory feedback. Our control system has similarities to the model reference adaptive control (MRAC) system presented in [30], [31], which incorporates a reference model of a motion sequence, inverts its dynamics and applies the resulting torques in a controlled manner to the input data.

The torques for the driving torque controller are computed using the inverse dynamics torque formulation

$$T(q, \dot{q}) = \tau(q^m, \dot{q}^m) + \frac{1}{2}\dot{q}^T \frac{\partial}{\partial \dot{q}} B \dot{q} - \dot{B}\dot{q}. \quad (14)$$

where τ is the vector of applied torques from the controller, and joint angles q^m and angular velocities \dot{q}^m are from the influencing gesture sequence. The joint configurations are transformed so that they correlate with the learned model's joint configurations.

Since there is no feedback in the driving torque controller, the torques can be precomputed. When $T(q, \dot{q})$ is applied to the dynamics it influences the motion of the model to follow a trajectory analogous to the influencing sequence. However, it is not necessarily strongly influencing the raw motion data to move towards the learned motion sequence. The strength of the influence is controlled by a scaling parameter k_c that is applied to the Kalman filter's process model error covariance matrix Q . This affects how much the system "trusts" the raw motion data versus the dynamic model. As k_c changes it directly impacts how the reported controller error relates to the measurement error in the system. As a result, the Kalman filter's gain matrix K (Equation 2), stabilizes differently, therefore changing how the Kalman filter weights input motion versus controller influence.

The blending function supplements the driving torque controller by providing more guidance to the state estimation. The driving torque controller provides the dynamics drive for the model, but it does not always provide sufficient guidance. The influencing motion sequence's torques may be nonlinear with respect to the

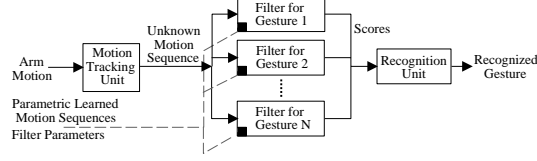


Fig. 4. Test Recognition System Architecture

joint configurations, but the tracking system performs blending of joint configurations linearly. Therefore, due to linear blending, small changes in the joint configurations can produce large changes in the dynamics. This directly affects how the driving torque controller performs. The blending function is intended to counteract this effect.

The blending function incorporates the current state of the system with the raw motion data from a learned motion sequence. The raw motion data includes the joint angles and angular velocities. This data is linear with respect to the motion state configurations of the system. The blending function that we use is

$$\mathbf{x}_{i+1} = b(\mathbf{x}_i + \Delta t \dot{\mathbf{x}}_i) + (1 - b)\mathbf{x}_i^m, \quad (15)$$

where $\mathbf{x}_i = [q, \dot{q}]^T$, $\dot{\mathbf{x}}_i = [\dot{q}, \ddot{q}]^T$, $\mathbf{x}_i^m = [q^m, \dot{q}^m]^T$, Δt is the current time step, and b is the blending factor.

5. Analysis of Filter

In order to test its effectiveness, we implemented our new filter, selected a difficult-to-discriminate gesture dataset, and ran user studies.

5.1. Design of Test System

We designed a system to test the motion adaptation filter by adapting a simple template-style gesture recognizer. We chose the template recognition system because it is easy to implement and is very easy to understand. However, our filter can work with most standard recognition architectures with some minor modifications (e.g., see notes in Section 7). The template architecture works by comparing the unknown input sequence with each gesture pattern. For our case, the unknown input is passed through a motion adaptation filter associated with each gesture (see Figure 4 for an overview).

Human motion data is brought into the system by a motion tracking unit and segmented by searching for long pauses in the motion sequences. The choice of tracking system is arbitrary, as long it can generate a continuous sequence of motion states. For this architecture, the output is distributed in parallel to N copies of the filter. Each of the filters is custom-tuned for a specific gesture. The output of the filters is a set of scores that are processed by the recognition unit. The scores are the squared differences of the internal unaugmented and augmented filters.

Although our filter can accept tracking data from any motion capturing technology, for purposes of testing we found it convenient to use a magnetic tracking system. There are obviously more accurate input technologies (e.g., acoustic

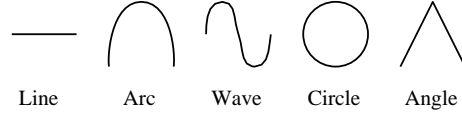


Fig. 5. Wrist-Trajectory Shapes of the Gesture Datasets used for the Expert User Experiments

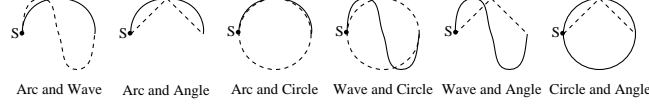


Fig. 6. Overlapping Features Embedded in Gesture Pairs

and inertial) and vision systems, but due to occlusion, they do not guarantee a continuous reliable stream of input.

We capture orientations of the lower arm, upper arm, and torso to retrieve the required four Euler angles. We estimate angular velocities using time difference methods. The set of angles and angular velocities makes up a motion state vector. The sequence of state vectors is sent to the motion state estimation unit.

5.2. Selection of a Hard-to-Discriminate Gesture Dataset

Our first step for analyzing the performance of the filter was to select a set of gestures that are hard to distinguish from each other. The selection criterion was determined by observing trajectories of the wrist for each gesture. The trajectories for the gesture dataset we selected for the introductory experiments are shown in Figure 5. This gesture set has many overlapping features, as can be seen in Figure 6. Two distinct gestures that have overlapping motion segments, especially if they start with the same motion sub-sequence, are more difficult to distinguish than dissimilar nonoverlapping gestures. A properly tuned EKF bases its initial output more on the input data than the dynamic model. But, when it converges to a stable blending state, the dynamics of the system takeover. If two gestures have similar starting trajectories and abruptly change after the dynamics become more dominant, the system will initially fail to discriminate between the two gestures because the derived dynamics of the system are similar. Eventually the mixture of the two dissimilar segments of the gestures will influence and change the system behavior.

For our experiments, we also considered the direction in which the motion was performed, thus expanding the five basic shapes to ten. We used combinations of the five basic shapes to generate gesture datasets and test the performance, generalizability and extensability of our approach in four of five expert-user experiments.

5.3. Filter Parameters

Our filter requires a set of parameters that must be predetermined and tuned for individual gestures. The EKF requires error covariance data for the measurement and control processes. The dynamics update requires measurements from the user's arm. Each control system requires a blending constant and a learned motion sequence.

Parameter Determination To compute the measurement error covariance we affixed three motion tracking receivers in the user workspace to a stationary configuration analogous to that of the right arm. We recorded 1000 samples continuously and computed the error covariance matrix computed using the sampled angles and estimated angular velocities. The measurement covariance matrix needs to be computed once for a given combination of hardware and workspace.

The control process error is computed by using the pre-recorded gesture sequences. A parametric learned motion sequence for each gesture type is selected by determining the closest fitting trajectory to a normal trajectory that is computed from the sample set of gestures. The error matrix is estimated using the mean squared error between the parametric learned motion sequence and the rest of the sequences. The control error needs to be computed for every gesture sequence.

Subject Measurements Some of the parameters needed for the filters are taken from measurements of the users. The filters require the lengths, radii and masses of the upper and lower arm. These parameters are obtained by combinations of two methods: direct measurements and estimation from anthropometric parameters of the human body. The lengths are determined by either directly measuring the distance between the shoulder and elbow, and elbow and wrist, or estimating them from the height and sex of the user. Estimations of anthropometric parameters are made according to the procedure outlined in Hall [29]. The radii are obtained by measuring the circumferences of the arm segments at the midpoint. The masses for the arm segments are determined as percentages of the whole body mass for males and females.

Parameter Tuning In order to use the EKF, specific parameters have to be tuned in order to get desirable guidance in the recognition units. One of the parameters that needs tuning is a multiplicative factor k_c used to scale the augmented filter’s control error covariance. There is one such scaling factor for each control error covariance matrix. The scaling factor is used to adjust the level of “trust” in the filter by changing the control error with respect to the measurement error. The larger k_c is, the more the filter output depends on the input. The smaller k_c is, the more the filter output depends on the controller and dynamic model. As a result the Kalman gain matrix, essential for the Kalman blend, changes. A similar single parameter is adjusted for the unaugmented filter.

Another parameter to be tuned is the blending factor b . This is applied in the blending function, which performs a blend of the intermediate state vector \mathbf{x}_{i+1}^{int} and the parametric learned motion sequence. This factor is important because it weights how the raw data is blended with the parametric learned motion sequence. The Kalman blend does not directly incorporate knowledge of the parametric learned motion sequence. We used one blending factor for all the gesture types. More details about the choice-of and tuning of these parameters is described in our technical report [4].

An important consideration when selecting the parameters is the degree of alignment of the input gesture with respect to the learned gesture. In the exper-

iments, we ask the users to extend their right arm perpendicular to the chest. The gestures they are asked to perform are then roughly centered around that hand position. Rough alignment and scaling is applied to the parametric learned gesture in addition to the parameterizing that is necessary to perform a matching comparison. This is the registration phase, which can be seen on the right side of the filter diagram in Figure 1. If the parametric learned gesture does not align very well with the gesture it is supposed to accept, it creates a high score for the comparison. This is due to our method for evaluation which compares the augmented and raw input trajectories. If the alignment is extremely bad we could not adjust the k_c parameter to “trust” the model as much. In most cases this is not a problem, but for a difficult dataset to recognize, such as the basic five gestures in Figure 5, some gestures will be improperly classified.

Sensitivity Analysis If we were to run a full user study on human subjects of widely varying mass and height, it would be important to understand how much of an impact parameter changes have on the dynamics of the system. If it can be shown that the system is relatively insensitive to changes in the parameters then it may be considered to be more generalizable and potentially more powerful. We analyzed the sensitivity of a few of the body parameters (summarized in Schmidt [32]), but did not determine enough meaningful information to make conclusions about the generalizability of our filter.






5.4. Expert User Experiments

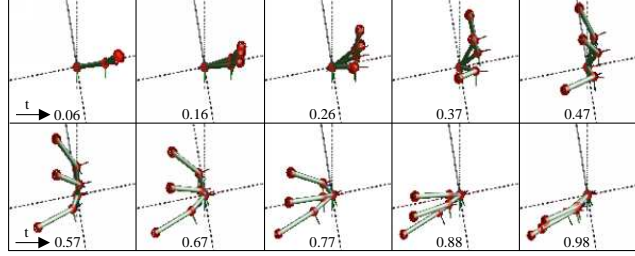
We set out to verify the effectiveness of the filter integrated into a gesture recognizer by devising a set of experiments to be performed by an expert user. These were designed to test the performance of the recognizer with and without our filter. We also wanted to ascertain something about how generalizable and extensible our filter is with respect to different and larger gesture datasets. To accomplish these goals, we ran five experiments. Before beginning, we pre-recorded a database of gestures from the user, computed the parameters and learned models, and performed manual parameter tuning.

Accuracy Performance The purpose of the first experiment was to determine the performance rating of the recognizer integrated with and without our filter. We used the five gestures from Table 1, and recorded 100 samples for each gesture. The gestures were first aligned with the learned motion sequences, then the learned motion sequences were parameterized to match the size of the input sequence. We supplied both the filtered (our method) and unfiltered recognizers with the 500 gestures. The results are given in Table 1.

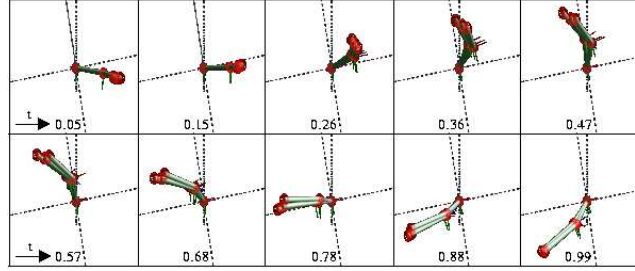
They show that both methods have an accuracy rating of 99.4%. The fact that both methods produced acceptable results turned out to be only coincidental for the unfiltered approach, which was later shown to be very inconsistent. We analyzed this dataset further and noticed that the gestures were fairly spatially regular with respect to each other. For example, there was not an extensive amount of variation due to alignment, skewing and scaling among the like gestures in this set.

Table 1. Results of Experiment #1

					Totals
Arc	Line	Wave	Circle	Angle	
99/100 99%	99/100 99%	100/100 100%	100/100 100%	99/100 99%	Unfiltered Approach 99.4%
98/100 98%	100/100 100%	100/100 100%	100/100 100%	99/100 99%	Our Filtered Approach 99.4%



a) Line in Arc Module



b) Arc in Arc Module






Fig. 7. Arm Model Motion in Time

To get a better idea of how our method works, refer back to Figure 2. The arc in the arc module shows the best match between the augmented and the unaugmented (effectively the learned motion sequence) trajectories. The rest of the cases show that the learned arc sequence has a large influence on the data running through the augmented filter which is evident by the output augmented trajectories. This effect pulls the augmented and raw data curves apart. The sequences in Figure 7 illustrate a small set of state transitions from the three arm models used in generating the trajectories for the line and the arc in the arc module. The figures show frames from a 3D simulation of the corresponding schematic 4-DOF arm models. The arm states are very similar for the arc in the arc module, but very different for the line in the arc module.

Generalizability To test the generalizability of our approach, we ran a second experiment. In the experiment we used the reverse-order wrist trajectories from the gestures used in the first experiment (a completely unique dataset). We recorded 100 samples for each of the five gestures and purposely added noise

into the samples to test the robustness of our filter. Then we passed them into the gesture recognizer twice, with and without our filter in the system. The resulting performance ratings are given in Table 2.

Table 2. Results of Experiment #2

 Arc	 Line	 Wave	 Circle	 Angle	Totals
60/100 60%	100/100 100%	78/100 78%	62/100 62%	99/100 99%	Unfiltered Approach 79.8%
98/100 98%	100/100 100%	100/100 100%	100/100 100%	96/100 96%	Our Filtered Approach 98.8%

In this case, the accuracy of the recognizer integrated with our filter proved to be far superior than without it. The performance rating for our filtered approach is 98.8%, while the unfiltered is 79.8%.

Extensibility For the third experiment, we examined the extensibility of our approach. To do this, we increased the number of distinct gestures that the recognizer had to distinguish. We used the two sets of gestures from the first two experiments and combine them into one database. Although diagrams make the two gesture sets appear similar, the motions that the human subject has to perform with the arm are totally different. When we performed the same experimental procedure as before, the results show our method has an accuracy rating of 99.1% while the unfiltered approach has a rating of 89.6%. This gives us a good indication that our method is extensible to larger size gesture datasets.






More Generalizability Experiments At this point we decided to revisit the first experiment with the hope of making it more difficult to distinguish the gestures than before. The goals of the fourth experiment were to show more generalizability with our method. In order to do this, we replaced the line and the wave with a triangle and another form of the arc. The new arc gesture is generated using a bend at the elbow instead of the straight arm motions used for the original arc. By our definition of arm gestures (i.e. movements of the arm that may or may not have any meaningful intent) and our analysis of only the “end-effector” position of the arm at the wrist, we do not make any distinction between the new and old arc gesture since both have identical wrist trajectories. The triangle gesture resembles the angle gesture in the first time steps, but deviates from it near the end. Our assumption was that this choice of gestures would be harder to discriminate. 75 trials were run for each gesture.

The experimental results show that the new gesture set was a bit harder to recognize by both methods. The triangle and bent-arm arc were recognized 90.7% and 86.7%, respectively for the unfiltered approach, and 98.7% and 96.0% for our approach. Our filtered approach showed an overall accuracy rating of 98.1% compared with the unfiltered approach’s rating of 95.2%. The results were again

encouraging with regard to our method’s consistency and accuracy, and also that it generalizes to different gestures quite well.

For our fifth experiment we ran 50 trials with five new gestures, each significantly different from the others. In addition, we decided to make a choice of somewhat natural gestures. The goal of the experiment was to determine if our method works well with gestures that are very easy to distinguish because they are quite distinct and are more natural. Our choices included the “zorro” sign, Catholic cross, salute, wave, and stop gestures. Diagrams of the motions of the wrist and results of the experiment are shown in Table 3.

Table 3. Results of Experiment #5

 Zorro	 Catholic Cross	 Waving	 Stop	 Salute	Totals
50/50 100%	46/50 92%	50/50 100%	50/50 100%	50/50 100%	Unfiltered Approach 98.4%
50/50 100%	50/50 100%	50/50 100%	50/50 100%	50/50 100%	Our Filtered Approach 100%

The results show that our method was 100% accurate on this gesture set, while the unfiltered approach achieved an accuracy rating of 98.4%.

Discussion In the experiments, we evaluated the accuracy performance, generalizability and extensibility of our filter when integrated in a recognition system. We made steps to ensure that it was difficult to distinguish among gestures by carefully selecting gesture datasets with overlapping motion traits. When compared with the recognizer with no filter attached, our method showed improved recognition performances. Our results from the five experiments show that our method is consistently accurate with rates ranging from 98.1% to 99.4% and extends to multiple gesture datasets. This compares very favorably with the unfiltered method whose accuracy ranged from 79.8% to 99.4%.

6. Pilot Study

We performed a pilot study involving six different subjects, in order to evaluate our model-based approach across different subjects.

6.1. Subject Selection

For the experiment, we selected three males and three females, with varying anatomical proportions. The sex discriminant was desired to accommodate for potential differing mass distributions in the arm between male and female subjects, based on muscle and bone proportions. The proportions we were concerned with were the lengths, radii and masses of the upper and lower right arm. The subjects were selected without regard to ethnicity, age, social or cultural backgrounds. The only screening requirement we had was a visual observation of size proportions in order to assure a subject pool of varying anatomical proportions.

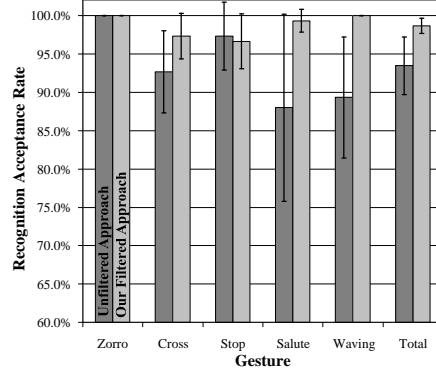


Fig. 8. Comparison of the Unfiltered and Filtered Approaches

6.2. Subject Measurements

The subjects had body weights ranging from 55 to 87 kg and heights ranging from 1.6 to 1.9 m, giving us a broad spectrum of masses and lengths for the user’s arm proportions. The upper arm lengths varied from 28 to 36 cm and the lower arm lengths from 23 to 27 cm. The upper arm radii varied from 3.66 to 5.25 cm and the lower arm radii from 3.18 to 4.38 cm. Trackers were attached using velcro straps at the wrist and near the elbow. A third was affixed with tape to the shoulder.

6.3. Pilot Experiment

In the first subject experiment our goal was to compare the difference between augmenting the recognition process with a model versus not augmenting the process. The subjects were asked to perform 25 trials of each of five different gestures, using the right arm. In between each set of trials for one gesture, the subject was given ample rest time to help avert any fatigue associated with the repetitive motions they were asked to make. We used the same five gestures as illustrated in Table 3, the zorro, Catholic cross, stop, salute and waving gestures.

The results we obtained were measurements of how well each recognition system predicted the correct gesture sequence. The performance rating for the two methods—the unaugmented and our model-based approach—were computed by averaging the performances for each of five different gestures. The performance for each gesture was computed by averaging the results from each of the six subjects. The histogram chart shown in Figure 8 compares the two sets of data.

The data for each user was analyzed by setting the body parameters for the recognizer to their measurements before running the accuracy tests. The rest of the parameters for the recognizer were individually tuned for each subject. The results for our model-based approach show an overall acceptance rate of 98.7% with standard deviation of 1.0%. The unaugmented approach performed at 93.5% acceptance rate with standard deviation of 3.7%. The high acceptance rate and low variability that our results show give us a fairly good indication that integrating our filter into the recognition process improves recognition accuracy.

A drawback of this experiment is that a significant amount of custom parameter tuning was required for each subject. As a result, we decided to evaluate whether or not our methodology would allow us to reduce the tuning effort required by each experiment. We ran a set of followup experiments to test these ideas. The results were somewhat limited. More details can be found in our technical report [4].

7. Discussion and Conclusions

We have developed a new model-based filter that incorporates a dynamics model, a control system and motion state estimation and applied it to the gesture recognition process. The dynamic model gives us a way to represent the underlying mechanical motion of the human arm. The control system acts as a means to exert control over and provide guidance for the motion applied by the dynamics.

Our filter proved to be effective in improving the performance of the recognition process as shown by our expert-user and pilot user studies. We showed this by comparing an unfiltered recognition process with one augmented with our model-based filter. Our method works acceptably well for hard-to-distinguish gesture sets and even better for very dissimilar sets. The results definitely warrant further user evaluation studies.

Our method does involve a small amount of parameter tuning and training for the error covariances. A lot of the tuning is associated with the registration of the input and learned gestures. Obviously, if the registration problem can be solved, a lot of the tuning can be eliminated. It also might be the case that more sophisticated models for the human motion or a more extensive model of the human body would reduce the need for some of the parameters.

One issue that our work did not address is the differences that may occur with people tracing the same “end-effector” path with different arm and joint configurations. For example, the “bent-arm” arc used in the fourth expert-user experiment has an equivalent wrist trajectory as the “straight-arm” arc had in the first experiments. We analyzed only the wrist trajectories, although we could have additionally analyzed either the elbow or joint configuration trajectories. This in effect increases the size of the gesture alphabet.

We only tested our filter with a template recognition architecture. However, we feel that it can be easily modified for use with a neural network recognizer. By removing the unaugmented sub-filter component the only output would be the augmented filtered sequence. If we setup n filters so that each input to the system produces n output sequences from the filters (for n distinct gesture patterns), each of these outputs will be different amongst themselves but fairly unique for each given input pattern. Then, extracting features from each output sequence which could yield $m \times n$ different features for the neural network. If desired, more features could be added from the raw or unaugmented filtered input sequence. The rest should follow the same as any neural network. The advantage of this (untested) setup would be that the filter could be used to generate many more unique discriminating features. While this is not always an advantage, if the

features are good discriminating ones we believe the discriminator should be more powerful.

Based on our evaluation studies, we can conclude that our motion adaptation filter makes a positive contribution to the performance of gesture recognition for arm-based gestures. This seems to imply that a model of human performance can be used to eliminate some of the heuristic guess-work that must be done to make a standard gesture recognizer work.

References

1. R. Watson, "A Survey of Gesture Recognition Techniques", Tech. Rep. TCD-CS-93-11, Department of Computer Science, Trinity College, Cambridge, U.K., July 1993.
2. J.K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review", in *IEEE Non-rigid and Articulated Motion Workshop 1997*, Piscataway, NJ, June 1997, IEEE.
3. V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–695, July 1997.
4. G. S. Schmidt and D. H. House, "Model-Based Motion Filtering for Improving Arm Gesture Recognition Performance", Tech. Rep. Technical Report, Virtual Reality Laboratory, Naval Research Laboratory, Washington, DC, Sept. 2003.
5. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Old Tappan, NJ: Prentice Hall PTR, 1993.
6. J. Martin, D. Hall, and J. L. Crowley, "Statistical Gesture Recognition through Modelling of Parameter Trajectories", in *Gesture Workshop '99: Gesture-Based Communication in Human-Computer Interaction*, Berlin, Germany, Mar. 1999, Springer-Verlag.
7. P. R. Cohen, D. McGee, S. L. Oviatt, L. Wu, J. Clow, R. King, S. Julier, and L. Rosenblum, "Multimodal Interactions for 2D and 3D Environments", *IEEE Computer Graphics and Applications*, vol. 4, pp. 10–13, July/August 1999.
8. A. D. Wexelblat, "A Feature-Based Approach to Continuous-Gesture Analysis", Master's thesis, Massachusetts Institute of Technology, May 1994.
9. D. Rubine, "Specifying Gestures by Example", in *ACM SIGGRAPH Computer Graphics Conference Proceedings*, Las Vegas, NV, Aug. 1991.
10. P. D. Gader, J. M. Keller, R. Krishnapuram, J.-H. Chiang, and M. A. Mohamed, "Neural and Fuzzy Methods in Handwriting Recognition", *IEEE Computer*, vol. 30, no. 2, pp. 79–86, Feb. 1997.
11. L. W. Campbell, D. A. Becker, A. Azarbayejani, A. F. Bobick, and A. Pentland, "Invariant Features for 3-D Gesture Recognition", in *Second International Workshop on Face and Gesture Recognition*, Killington, VT, Oct. 1996.
12. G. Herzog and K. Rohr, "Integrating Vision and Language: Towards Automatic Description of Human Movements", in *Proceedings of the 19th Annual German Conference on Artificial Intelligence, KI-95*, Bielefeld, Germany, July 1995.
13. A. D. Wilson and A. Bobick, "Using Configuration States for the Representation and Recognition of Gesture", Tech. Rep. Technical Report No. 308, M.I.T. Media Laboratory, Cambridge, MA, June 1995.
14. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, New York, NY: John Wiley & Sons, Inc., 1973.

15. L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
16. M. T. Vo and A. Waibel, "A Multi-Modal Human-Computer Interface: Combination of Gesture and Speech Recognition", in *Adjunct Proceedings of InterCHI '93*, Apr. 1993.
17. A. Pentland and B. Horowitz, "Recovery of Nonrigid Motion and Structure", *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 13, no. 7, pp. 730–742, 1991.
18. V. B. Zordan and J. K. Hodgins, *Tracking and Modifying Upper-Body Human Motion Data with Dynamic Simulation*, pp. 13–22, Vienna, Austria: Springer-Verlag, Sept. 1999.
19. D. Metaxas, "Articulated Figure Dynamics, Behavior and Control", in *Virtual Humans: Behaviors and Physics, Acting, and Reacting (SIGGRAPH '97 Course Notes)*, Los Angeles, CA, Aug. 1997.
20. C. R. Wren and A. P. Pentland, "Dynamic Models of Human Motion", in *Third IEEE International Workshop on Automatic Face and Gesture Recognition*, Nara, Japan, Apr. 1998.
21. C. R. Wren, B. P. Clarkson, and A. P. Pentland, "Understanding Purposeful Human Motion", in *4th International Workshop on Automatic Face and Gesture Recognition*, Grenoble, France, Mar. 2000.
22. G. S. Schmidt and D. H. House, "Towards Model-Based Gesture Recognition", in *4th International Workshop on Automatic Face and Gesture Recognition*, Grenoble, France, Mar. 2000.
23. G. Welch and G. Bishop, "An Introduction to the Kalman Filter", Tech. Rep. TR 95-041, Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, Dec. 1995.
24. R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems", *Transactions of the ASME—Journal of Basic Engineering*, vol. 1, no. 2, pp. 34–45, 1960.
25. R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, New York, NY: CRC Press LLC, 1993.
26. R. Featherstone, "The Calculation of Robot Dynamics Using Articulated-Body Inertias", *International Journal of Robotics Research*, vol. 2, no. 1, pp. 13–30, 1983.
27. J. J. Craig, *Introduction to Robotics: Mechanics and Control*, Reading, MA: Addison-Wesley Publishing Company, Inc., 1989.
28. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, New York, NY: Cambridge University Press, 1992.
29. S. J. Hall, *Basic Biomechanics*, St. Louis: Mosby, 1995.
30. I. D. Landau, *Adaptive Control, The Model Reference Approach*, New York, NY: Marcel Dekker, 1979.
31. D. P. Stoten and H. Benchoubane, "Empirical Studies of an MRAC Algorithm with Minimal Control Synthesis", *International Journal of Control*, vol. 51, no. 4, pp. 823–849, 1990.
32. G. S. Schmidt, *Model-Based Gesture Recognition*, PhD thesis, Texas A&M University, Computer Science Dept., Texas A&M University, Dec. 2000.